Fashion Retail Dataset - Exploratory Data Analysis Report

Generated: February 13, 2026 at 03:31 PM

1. EXECUTIVE SUMMARY

This report presents an exploratory data analysis of the H&M Personalized Fashion Retail dataset. The dataset includes product metadata, customer demographics, and transaction records. The goal of this analysis is to assess data quality, understand distribution patterns, and prepare the dataset for demand prediction modeling.

2. Dataset Overview

Metric	Value
Total Articles	105,542
Total Customers	1,371,980
Total Transactions (Sample)	50,000
Product Features	25
Customer Features	7
Transaction Features	5

The analysis focuses on:

- Data completeness
- Numeric feature behavior
- Distribution characteristics
- Outlier detection
- Categorical insights
- Demand behavior patterns

The dataset is suitable for large-scale demand forecasting with appropriate preprocessing.

3. Dataset Overview

Metric	Value
Total Articles	105,542
Total Customers	1,371,980
Transactions (Sample)	50,000
Weekly Demand Records	11,726
Article Features	25
Customer Features	7
Transaction Features	7

Interpretation

- The dataset is large-scale and structurally rich.
- The number of products is substantial, enabling product-level modeling.

Weekly demand records allow time-based demand modeling.

- The dataset structure supports supervised learning tasks.

4. Missing Values Analysis

Column	Missing Count	Missing %
articles.detail_desc	416	0.39
customers.FN	895050	65.24
customers.Active	907576	66.15
customers.club_member_status	6062	0.44
customers.fashion_news_frequency	16011	1.17
customers.age	15861	1.16

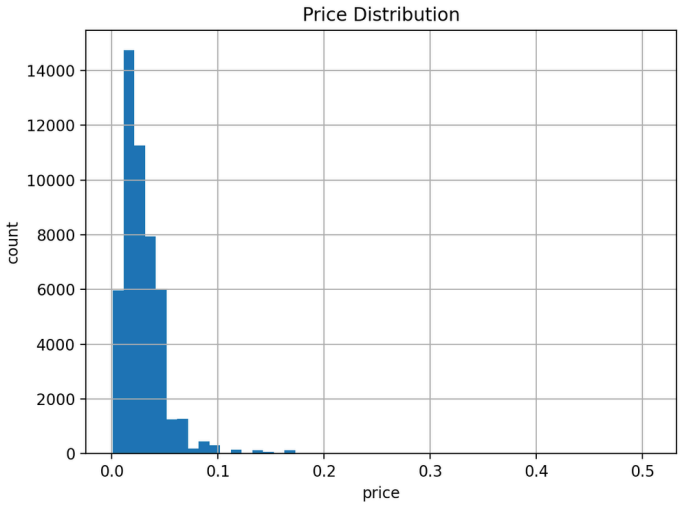
Interpretation

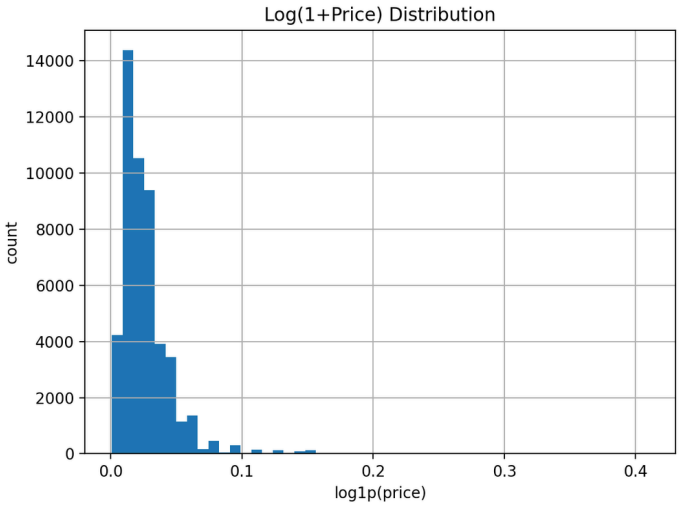
- Customer behavioral flags (FN, Active) are heavily incomplete.
- These variables may need:
  - Imputation
  - Encoding as "Unknown"
  - Or exclusion depending on modeling relevance.
- Transaction data is clean and reliable, which is critical since it defines demand.

5. Numeric Features Statistics

Feature	Mean	Std	Min	Max	Skewness
transactions.price	0.02926	0.02222	0.00085	0.50678	3.76109
weekly_demand.demand	4.26403	9.12128	1	455	18.14552
customers.age	36.38696	14.31363	16.0	99.0	0.61259

5.1 Price Analysis



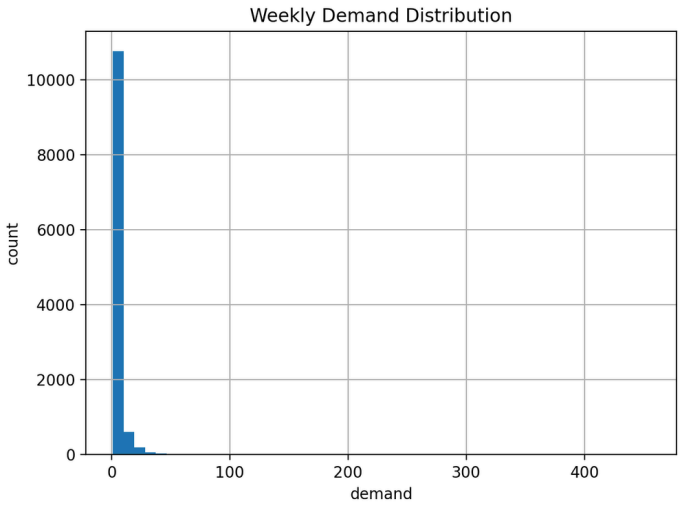


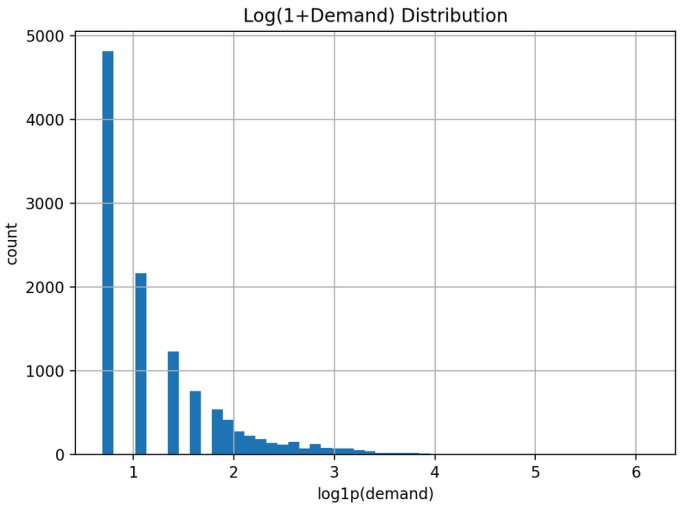
- Strong positive skewness (3.76)
- Most products are low-priced
- Few high-price products create a long tail
- Majority of transactions occur at lower price range
- Distribution is heavily right-skewed

Interpretation

Retail sales are concentrated in affordable price segments, suggesting high turnover in lower-price items.

5.2 Weekly Demand Analysis





Skewness = 18.14 (extremely high)

- Most products have low weekly demand
- Very few products have extremely high demand
- Long right tail visible

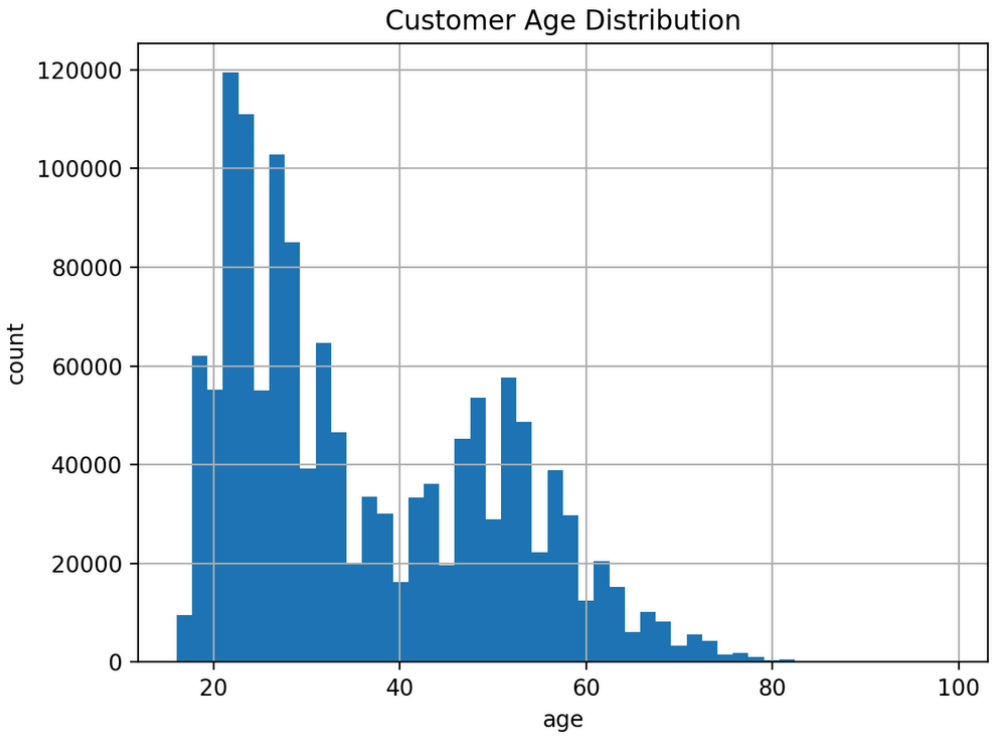
Interpretation

Demand is highly imbalanced.

This suggests:

- Popularity follows a power-law distribution
- Few products drive most transactions
- Demand prediction will need:
  - Log transformation
  - Or classification thresholds

6. Customer Age



  - Mean age ≈ 36 years
  - Slight right skew (0.61)
  - Majority between 20–45 years
  - Smaller segment 50+
- Interpretation
- Primary customer segment: young to middle-aged adults.

7. Distribution Analysis

Log Transformations

Log(1+price) and Log(1+demand) plots show:

- Reduced skewness
- More normal-like distribution

- Better modeling suitability

Modeling Implication

Log transformation recommended for:

- Demand regression
- Price-based features

6. Outlier Detection (IQR Method)

Feature	Outlier Count	Outlier %
transactions.price	3778	7.56
weekly_demand.demand	1294	11.04
customers.age	214	0.02

Interpretation

- Demand contains notable extreme values.
- High-demand products are statistical outliers.
- Age is very stable.

Business Meaning

Outlier demand products may represent:

- Viral items
- Seasonal peaks
- Promotional effects

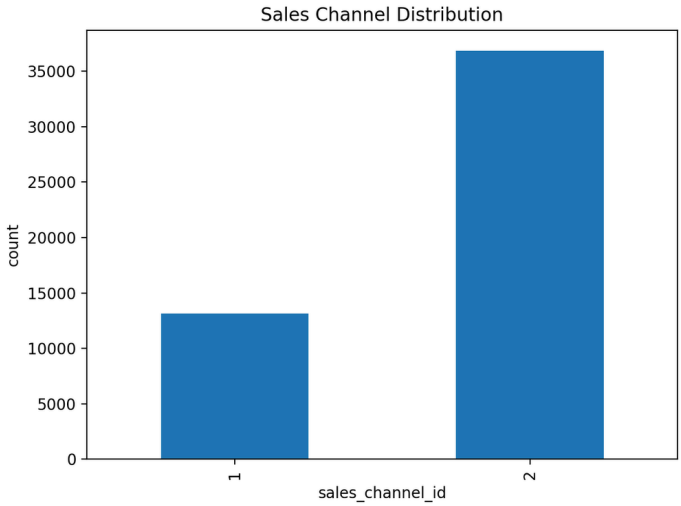
Decision needed:

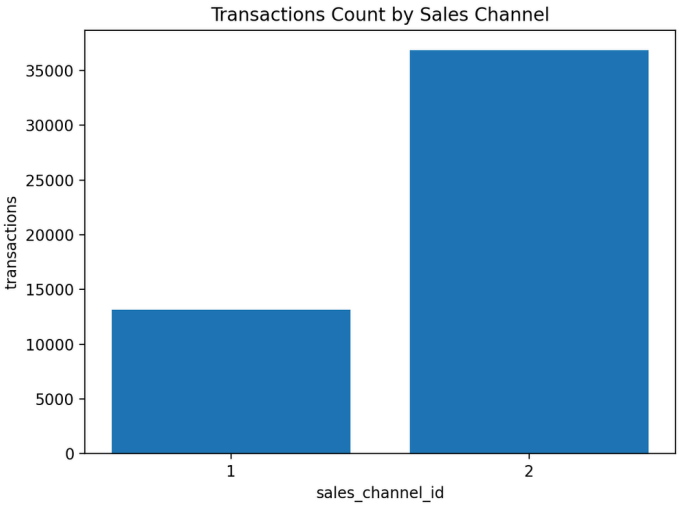
- Keep outliers (business-critical)

Or cap them (robust modeling)

7. Categorical Distributions

7.1 Sales Channel



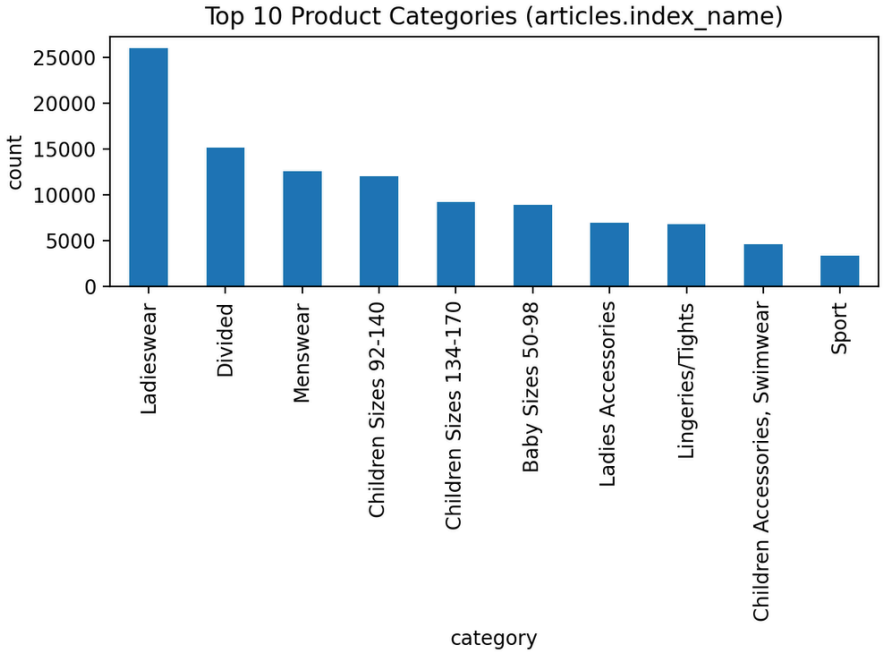


- Channel 2 dominates transactions.
- Channel 1 significantly lower volume.

Interpretation

Channel 2 likely represents online sales, indicating digital dominance.

7.2 Product Categories



Top categories:

- Ladieswear
- Divided
- Menswear
- Children Sizes

Interpretation

- Women's fashion dominates product inventory.

- Children and men's categories also significant.
- Demand modeling may benefit from category-level features.

8. Key Analytical Findings

1. Demand distribution is highly skewed.
2. A small number of products generate large volume.
3. Retail transactions are price-sensitive.
4. Customer demographic distribution is centered around 20–45 years.
5. Online channel dominates sales activity.
6. Dataset is suitable for predictive modeling after preprocessing.

9. Modeling Readiness Assessment

Clean transaction data
Sufficient scale
Weekly aggregation available
Target variable defined
Outliers identified
Transformations explored

Next logical steps:

- Merge article metadata with weekly demand
- Encode categorical features
- Log-transform demand
- Train regression model
- Track with MLflow

10. Conclusion

The dataset is:

- Large-scale
- Structurally rich
- Suitable for supervised learning
- Representative of real retail dynamics

Demand shows strong imbalance and heavy skewness, requiring transformation and careful modeling strategy.